Dr. Eick

COSC 6335 *“Data Mining”* Fall 2023

Task 2: An Intelligent Tool Which Compares Boxplots

 

Task2 Due: Saturday, October 7 end of the day (electronic Submission via MS Teams)

Remark: Kritik will not be used for this task!

Last Updated: October 5, noon

**Learning Objectives**:

1. Comparing box plots
2. Converting numbers into natural language summaries
3. Tools for data storytelling

Develop a tool that compares 2 box plots and reports its findings as a story. The story summarizes the similarity of two input box plots[[1]](#footnote-1) with respect to median, box location, IQR, maximum and minimum values that are no outlier (and maybe[[2]](#footnote-2) also compares the reported outliers) and skewness; moreover, the tool summarizes the overall agreement / disagreement of the two boxplots. Develop an interface that that our TA Mahin, can run your tool; instructions how to run your program is part of the Task2 deliverables; moreover, a benchmark of 4 example pairs of boxplots to compare can be found at the end of this specification.

Basically, the input of the tool are the following values of each boxplot:

Median1, median2:= median of each boxplot

75perc1,75perc2:= 75% percentile value of each boxplot

25perc1,25perc2:= 25% percentile value of each boxplot

max1, max2:= maximum value in each dataset from which the box plot was created which is not an outlier

min1, min2:= minimum value in each dataset from which the box plot was created which is not an outlier

These values can be obtained for a boxplot b in R using: b.$stats

Other values that are derived from those input parameters:

IQR1, IQR2 := IQRs of the respective boxplots

If your tool additionally analyzes outliers, the following inputs are additionally considered:

Out1:= set of outlier values boxplot1; could be empty!

Out2:= set of outlier values boxplot2; could be empty!

For a boxplot b outliers can be obtained in R using: b$out

Based on those input values the tool creates a “summary story” concerning the similarity of the two input boxplots.

Four Test cases for Problemset1 Task2:

Please run the tool you develop for the following 4 pairs of box plots which have been produced from the listed 1D-datasets below:

**Case 1:**

L1 = [18, 58, 22, 50, 44, 64, 68, 10, 58, 6, 82, 42, 39, 26, 18, 44, 80, 26, 59, 35, 20, 81, 23, 77, 18, 72, 24, 3, 30, 81, 189, 149, 109, 139]

L2 = [58, 98, 90, 93, 55, 14, 15, 94, 33, 84, 26, 29, 98, 22, 24, 47, 52, 51, 71, 48, 83, 15, 0, 50, 99, 20, 21, 85, 57, 49]

**Case 2:**

class1 = [18, 58, 22, 50, 44, 64, 68, 10, 58, 6, 82, 42, 39, 26, 18, 44, 80, 26, 59, 35]

class2 = [58, 98, 90, 93, 55, 14, 15, 94, 33, 84, 26, 29, 98, 22, 24, 47, 52, 51, 71, 48]

**Case 3:**

Dallas = [18, 58, 22, 50, 44, 64, 14, 15, 58, 6, 26, 29, 26, 18, 44, 178, 170]

Houston = [149, 180, 165, 166, 147, 171, 154, 125, 136, 162, 141, 139, 131, 132, 143, 11, 34, 18]

**Case 4:**

Dallas = [18, 58, 22, 50, 44, 64, 14, 15, 58, 6, 26, 29, 26, 18, 44, 178, 170]

Tuscon = [19, 59, 23, 51, 45, 65, 15, 16, 59, 24, 25, 28, 25, 17, 43, 0, 1, 2]

Sample Comparison for Case1

Remark: We expect your tool only to output what is depicted in **brown** below. The other output you find below is optional. This is only an example to give you a better idea what your tool is expected to do and there are many other ways to compare the two boxplots!

**Case 1:**

L1 = [18, 58, 22, 50, 44, 64, 68, 10, 58, 6, 82, 42, 39, 26, 18, 44, 80, 26, 59, 35, 20, 81, 23, 77, 18, 72, 24, 3, 30, 81, 109, 149, 109, 139]

L2 = [58, 98, 90, 93, 55, 14, 15, 94, 33, 84, 26, 29, 98, 22, 24, 47, 52, 51, 71, 48, 83, 15, 0, 50, 99, 20, 21, 85, 57, 49]



Summary:

The medians of L1 and L2 are somewhat similar since the Median of L2 is 0.17\*box\_size higher than the median of L1. Comparing the IQR’s of the two distributions we found, the overlap of two IQRs is 67.76%, which indicates the IQR to be somewhat overlapping. The distribution of L1 and L2 have 79.80% overlap in whisker which indicates them to be quite similar distribution. We also observe both boxplots have almost no skew since the median is at the middle of the box. There is one outlier for L1 distribution but none in the L2. In summary, we conclude that the two boxplots are somewhat similar.

Submission instructions Task2:

1. Your code: Submit the code of the boxplot comparison tool. Comment on your code where appropriate. If you code in jupyter notebook, please submit the .py or .ipynb version.
2. Include instructions how to run your submitted program e.g. how to give input or expect output based on application type (console/GUI) application.
3. Discussion (pdf format): Provide a high level summary how your program compares the two boxplots. Include the output your program produces for the 3 test cases 2-4 in your report.

Put your code and discussion a zipped folder and submit it via MS Teams, folder assigned for the task.

1. Moreover, we assume that boxplots are generated using the default parameter setting of the respective boxplot function of the boxplot tool you use. [↑](#footnote-ref-1)
2. A small amount of extra credit (up to 5%) will be given to students whose storytelling tool compares the presence and location of outliers in the two box plots. [↑](#footnote-ref-2)